

Learning Sequence Motif Models Using Expectation Maximization (EM)

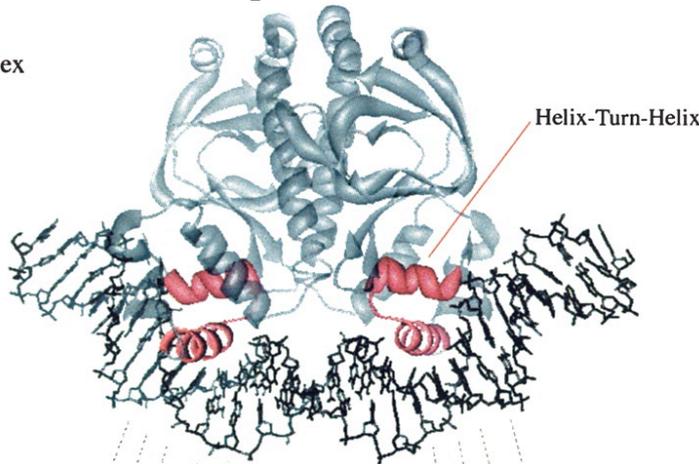
Colin Dewey
(adapted from slides by Mark Craven)
2007.02.06

Sequence Motifs

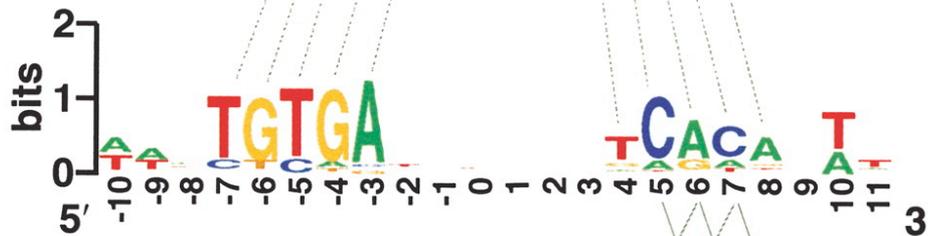
- what is a sequence *motif*?
 - a sequence pattern of biological significance
 - typically repeated several times in the genome
- examples
 - protein binding sites in DNA
 - protein sequences corresponding to common functions or conserved pieces of structure

Sequence Motifs Example

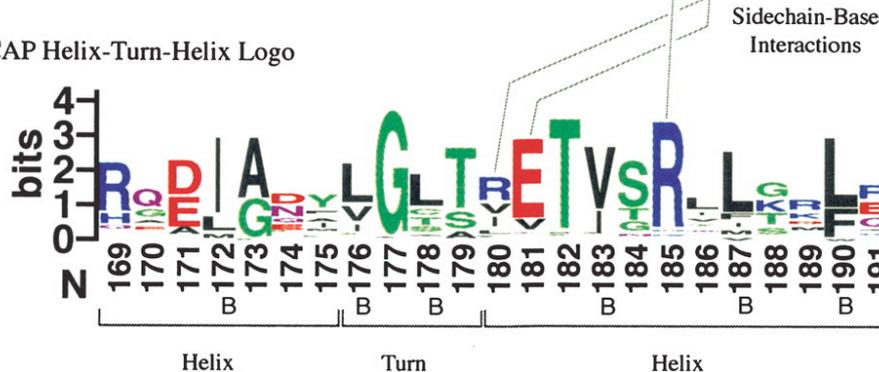
A CAP-DNA Complex



B CAP recognition site DNA Logo



C CAP Helix-Turn-Helix Logo

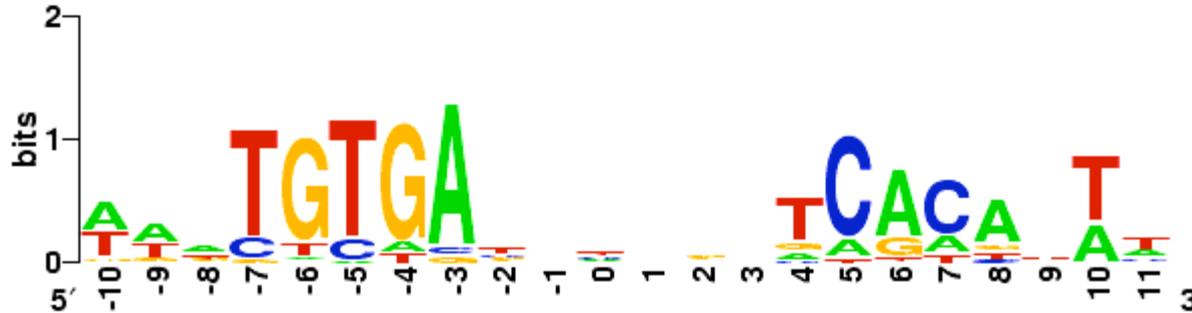


CAP-binding motif
model based on 59
binding sites in E.coli

helix-turn-helix motif
model based on 100
aligned protein
sequences

Figure from Crooks et al., *Genome Research* 14:1188-90, 2004.

Sequence Logos



- height of the scale is determined by the maximum *entropy* at a sequence position: $H_{\max} = \log_2 N$ where N is the number of possible characters
- height of logo at a given position is determined by decrease in entropy at that position

$$H_{\max} - H_{\text{obs}} = \log_2 N - \left(- \sum_{n=1}^N p_n \log_2 p_n \right)$$

where p_n is the frequency of character n at the position

The Motif Model Learning Task

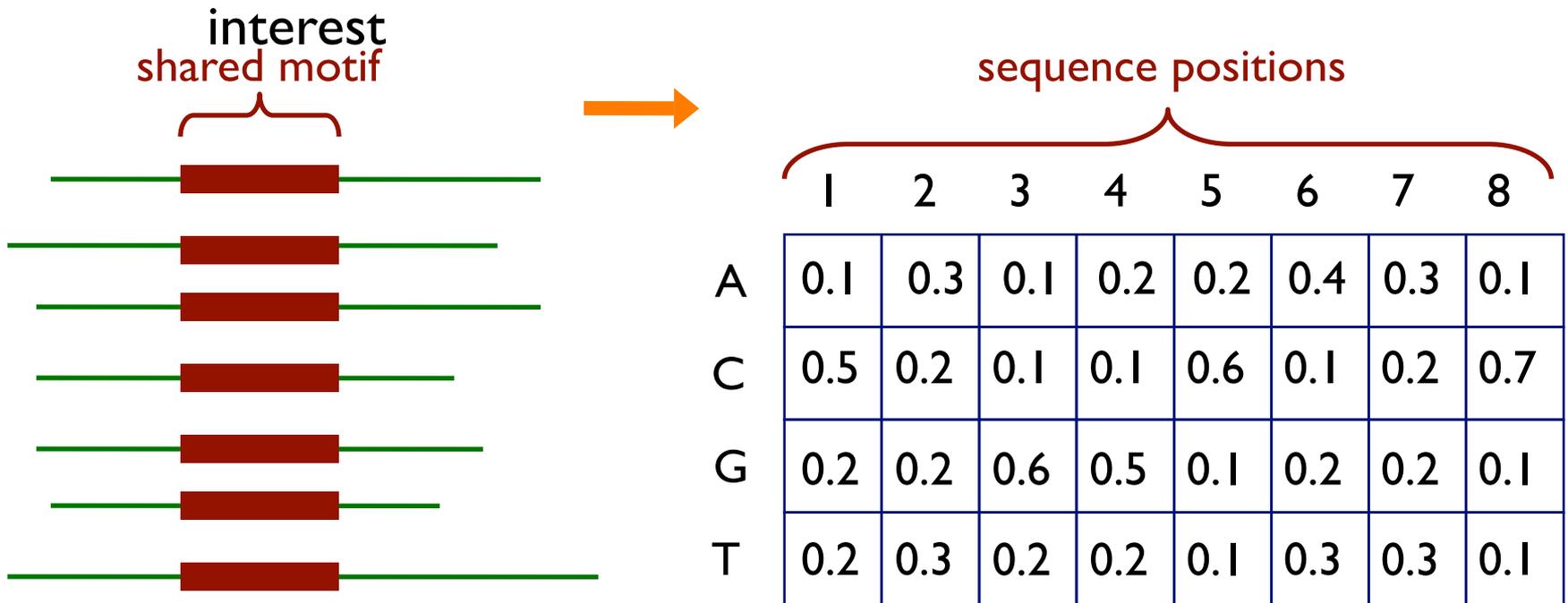
given: a set of sequences that are thought to contain an unknown motif of interest

do:

- infer a model of the motif
- predict the locations of the motif in the given sequences

Motifs and *Profile Matrices* (a.k.a. *Position Weight Matrices*)

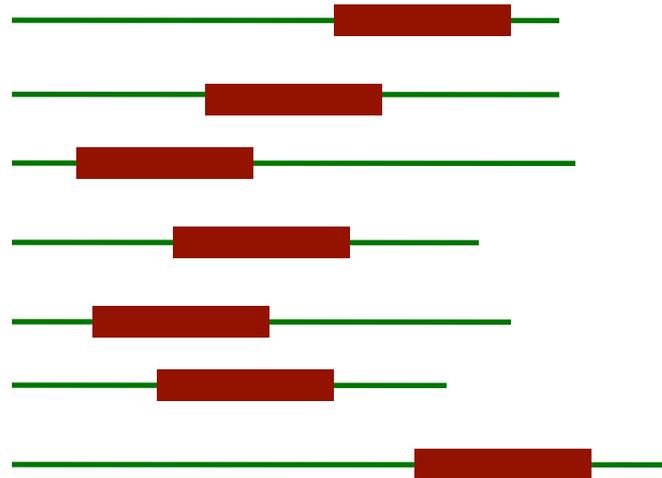
- given a set of aligned sequences, it is straightforward to construct a profile matrix characterizing a motif of



- each element represents the probability of given character at a specified position

Motifs and Profile Matrices

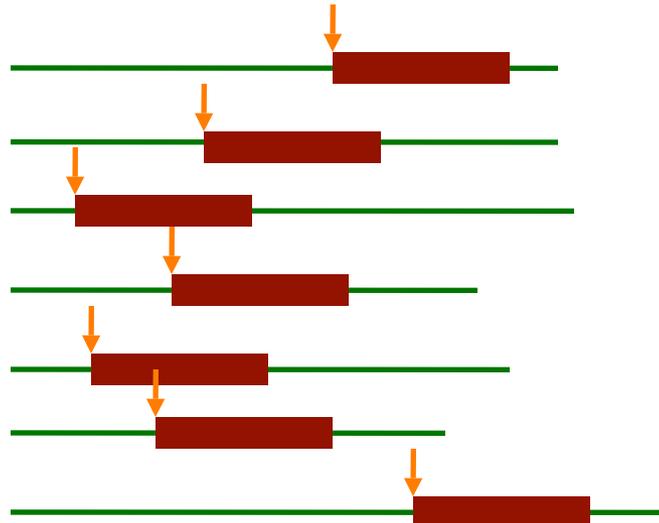
- how can we construct the profile if the sequences aren't aligned?
- in the typical case we don't know what the motif looks like



- use an Expectation Maximization (EM) algorithm

The EM Approach

- EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*
- in our problem, the hidden state is where the motif starts in each training sequence



The MEME Algorithm

- Bailey & Elkan, 1993, 1994, 1995
- uses EM algorithm to find multiple motifs in a set of sequences
- first EM approach to motif discovery:
Lawrence & Reilly 1990

Representing Motifs in MEME

- a motif is
 - assumed to have a fixed width, W
 - represented by a matrix of probabilities: $p_{c,k}$ represents the probability of character c in column k
- also represent the “background” (i.e. outside the motif) probability of each character: $p_{c,0}$ represents the probability of character c in the background
- example: a motif model of length 3

		0	1	2	3
$p =$	A	0.25	0.1	0.5	0.2
	C	0.25	0.4	0.2	0.1
	G	0.25	0.3	0.1	0.6
	T	0.25	0.2	0.2	0.1

Basic EM Approach

- the element $Z_{i,j}$ of the matrix Z represents the probability that the motif starts in position j in sequence i
- example: given 4 DNA sequences of length 6, where $W=3$

			1	2	3	4
	seq1	0.1	0.1	0.2	0.6	
$Z =$	seq2	0.4	0.2	0.1	0.3	
	seq3	0.3	0.1	0.5	0.1	
	seq4	0.1	0.5	0.1	0.3	

Basic EM Approach

given: length parameter W , training set of sequences

set initial values for p

do

re-estimate Z from p (E-step)

re-estimate p from Z (M-step)

until change in $p < \epsilon$

return: p, Z

Calculating the Probability of a Sequence Given a Hypothesized Starting Position



$$\Pr(X_i \mid Z_{i,j} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^L p_{c_k,0}}_{\text{after motif}}$$

X_i is the i th sequence

$Z_{i,j}$ is 1 if motif starts at position j in sequence i

c_k is the character at position k in sequence i

Example

$X_i = \text{G C } \boxed{\text{T G T}} \text{ A G}$

		0	1	2	3
	A	0.25	0.1	0.5	0.2
$p =$	C	0.25	0.4	0.2	0.1
	G	0.25	0.3	0.1	0.6
	T	0.25	0.2	0.2	0.1

$$\Pr(X_i | Z_{i3} = 1, p) =$$

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$
$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

The E-step: Estimating Z

- to estimate the starting positions in Z at step t

$$Z_{i,j}^{(t)} = \frac{\Pr(X_i | Z_{i,j} = 1, p^{(t)}) \Pr(Z_{i,j} = 1)}{\sum_{k=1}^{L-W+1} \Pr(X_i | Z_{i,k} = 1, p^{(t)}) \Pr(Z_{i,k} = 1)}$$

- this comes from Bayes' rule applied to

$$\Pr(Z_{i,j} = 1 | X_i, p^{(t)})$$

The E-step: Estimating Z

- assume that it is equally likely that the motif will start in any position

$$Z_{i,j}^{(t)} = \frac{\Pr(X_i | Z_{i,j} = 1, p^{(t)}) \cancel{\Pr(Z_{i,j} = 1)}}{\sum_{k=1}^{L-W+1} \Pr(X_i | Z_{i,k} = 1, p^{(t)}) \cancel{\Pr(Z_{i,k} = 1)}}$$

Example: Estimating Z

$X_i = \text{G C T G T A G}$

		0	1	2	3
$p =$	A	0.25	0.1	0.5	0.2
	C	0.25	0.4	0.2	0.1
	G	0.25	0.3	0.1	0.6
	T	0.25	0.2	0.2	0.1

$$Z_{i,1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i,2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

⋮

- then normalize so that $\sum_{j=1}^{L-W+1} Z_{i,j} = 1$

The M-step: Estimating p

- recall $P_{c,k}$ represents the probability of character c in position k ; values for $k=0$ represent the background

$$P_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1} = c\}} Z_{i,j} & k > 0 \\ n_c - \sum_{j=1}^W n_{c,j} & k = 0 \end{cases}$$

total # of c's in data set

Example: Estimating p

A C A G C A

$$Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1$$

A G G C A G

$$Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4$$

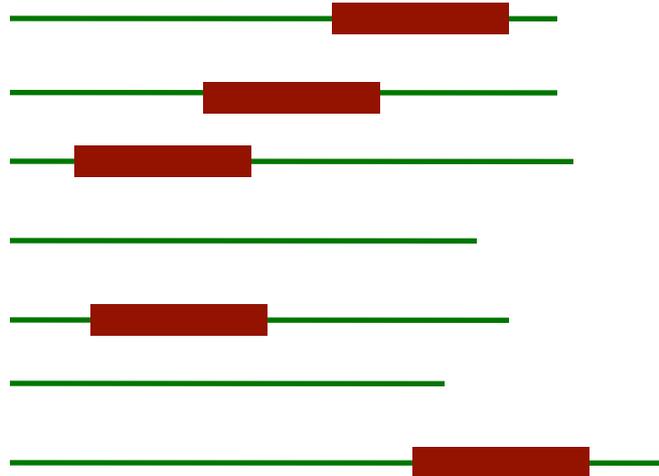
T C A G T C

$$Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1$$

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \dots + Z_{3,3} + Z_{3,4} + 4}$$

The ZOOPS Model

- the approach as we've outlined it, assumes that each sequence has exactly one motif occurrence per sequence; this is the OOPS model
- the ZOOPS model assumes zero or one occurrences per sequences



E-step in the ZOOPS Model

- we need to consider another alternative: the i th sequence doesn't contain the motif
- we add another parameter (and its relative)
 - λ prior probability that any position in a sequence is the start of a motif
 - $\gamma = (L - W + 1)\lambda$ prior probability of a sequence containing a motif

E-step in the ZOOPS Model

$$Z_{i,j}^{(t)} = \frac{\Pr(X_i | Z_{i,j} = 1, p^{(t)})\lambda^{(t)}}{\Pr(X_i | Q_i = 0, p^{(t)})(1 - \gamma^{(t)}) + \sum_{k=1}^{L-W+1} \Pr(X_i | Z_{i,k} = 1, p^{(t)})\lambda^{(t)}}$$

- Q_i is a random variable for which $Q_i = 1$ if sequence X_i contains a motif, $Q_i = 0$ otherwise

$$\Pr(Q_i = 1) = \sum_{j=1}^{L-W+1} Z_{i,j}$$

$$\Pr(X_i | Q_i = 0, p) = \prod_{j=1}^L p_{c_j,0}$$

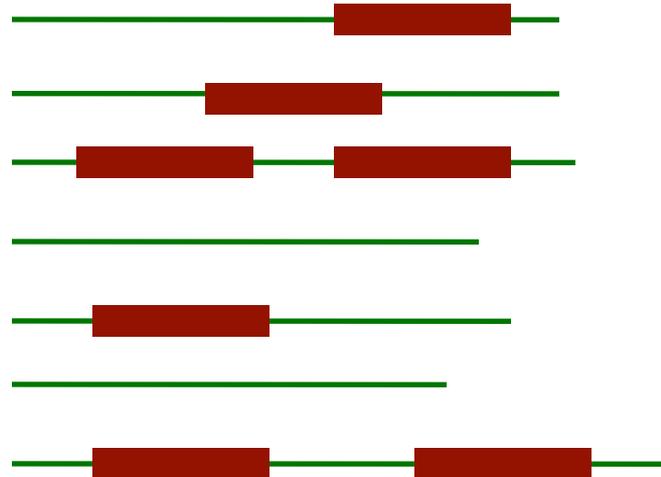
M-step in the ZOOPS Model

- update p same as before
- update γ as follows:

$$\gamma^{(t+1)} \equiv \lambda^{(t+1)} (L - W + 1) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{L-W+1} Z_{i,j}^{(t)}$$

The TCM Model

- the TCM (two-component mixture model) assumes *zero or more* motif occurrences per sequence



Likelihood in the TCM Model

- the TCM model treats each length W subsequence independently
- to determine the likelihood of such a subsequence:

$$\Pr(X_{i,j} \mid Z_{i,j} = 1, p) = \prod_{k=j}^{j+W-1} p_{c_k, k-j+1} \quad \text{assuming a motif starts there}$$

$$\Pr(X_{i,j} \mid Z_{i,j} = 0, p) = \prod_{k=j}^{j+W-1} p_{c_k, 0} \quad \text{assuming a motif doesn't start there}$$

E-step in the TCM Model

$$Z_{i,j}^{(t)} = \frac{\Pr(X_{i,j} | Z_{i,j} = 1, p^{(t)})\lambda^{(t)}}{\underbrace{\Pr(X_{i,j} | Z_{i,j} = 0, p^{(t)})(1 - \lambda^{(t)})}_{\text{subsequence isn't a motif}} + \underbrace{\Pr(X_{i,j} | Z_{i,j} = 1, p^{(t)})\lambda^{(t)}}_{\text{subsequence is a motif}}}$$

- M-step same as before

Extending the Basic EM Approach in MEME

- How to choose the width of the motif?
- How to find multiple motifs in a group of sequences?
- How to choose good starting points for the parameters?
- How to use background knowledge to bias the parameters?

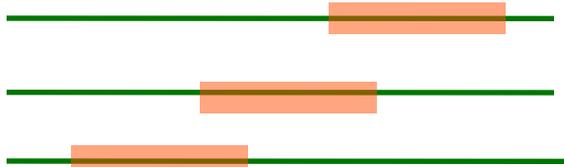
Choosing the Width of the Motif

- try various widths
 - estimate the parameters each time
 - apply a likelihood ratio test based on
 - probability of data under motif model
 - probability of data under *null* model
 - penalized by # of parameters in the model

Finding Multiple Motifs

- we might want to find multiple motifs in a given set of sequences
- how can we do this without
 - rediscovering previously learned motifs

iteration 1

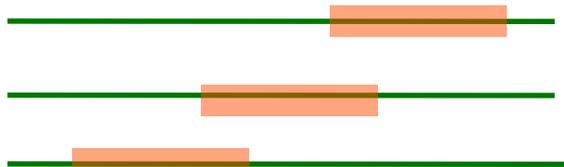


iteration 2

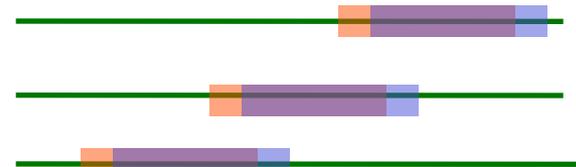


- discovering a motif that substantially overlaps with previously learned motifs

iteration 1



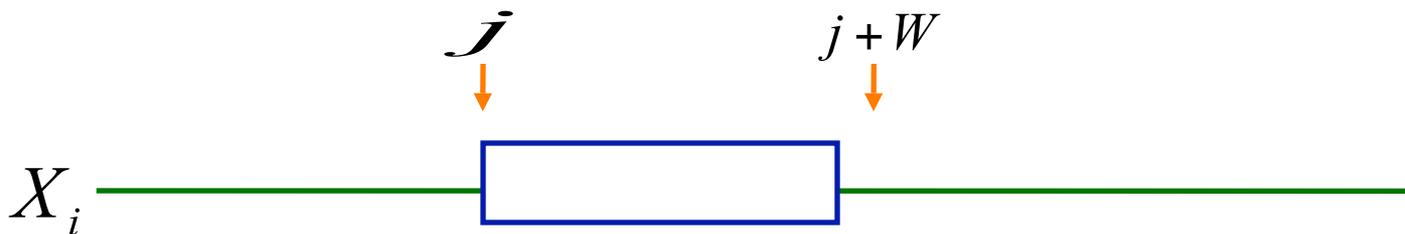
iteration 2



Finding Multiple Motifs

- basic idea: discount the likelihood that a new motif starts in a given position if this motif would overlap with a previously learned one
- when re-estimating $Z_{i,j}$, multiply by $\Pr(V_{i,j} = 1)$

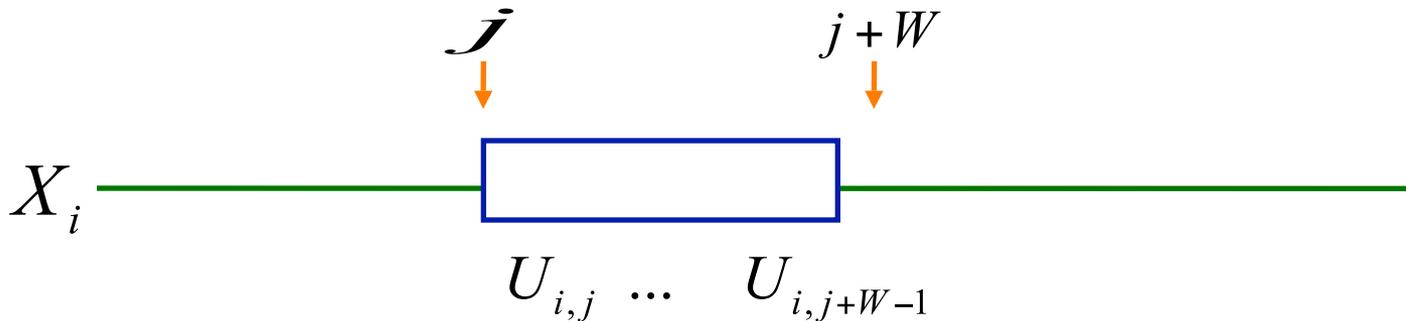
$$V_{i,j} = \begin{cases} 1, & \text{no previous motifs in } [X_{i,j}, \dots, X_{i,j+w-1}] \\ 0, & \text{otherwise} \end{cases}$$



Finding Multiple Motifs

- to determine $\Pr(V_{i,j} = 1)$ need to take into account individual positions in the window

$$U_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \notin \text{previous motif occurrence} \\ 0, & \text{otherwise} \end{cases}$$

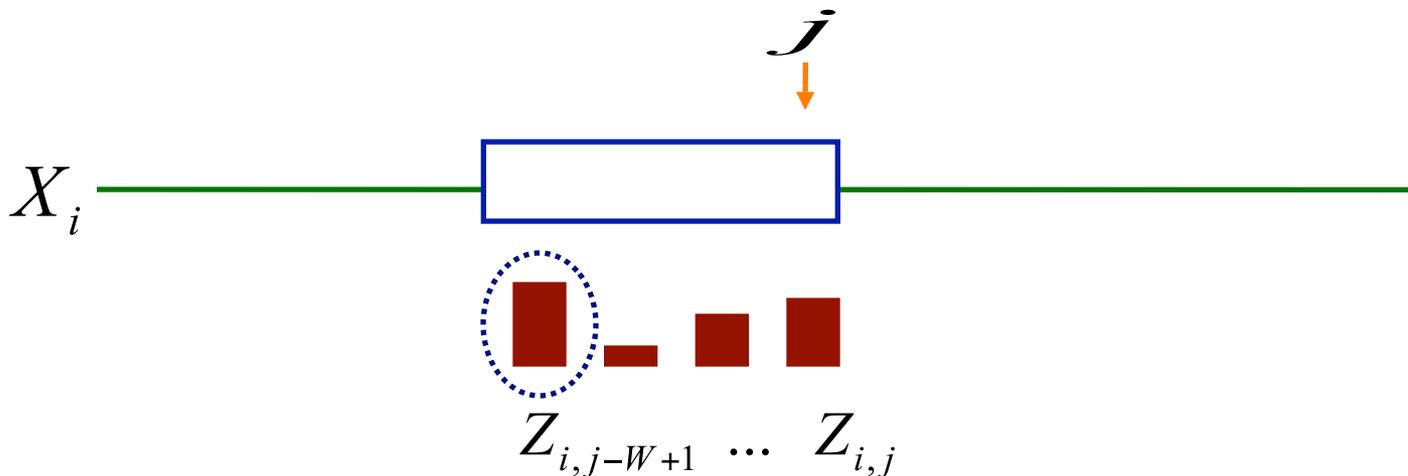


Finding Multiple Motifs

$$U_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \notin \text{previous motif occurrence} \\ 0, & \text{otherwise} \end{cases}$$

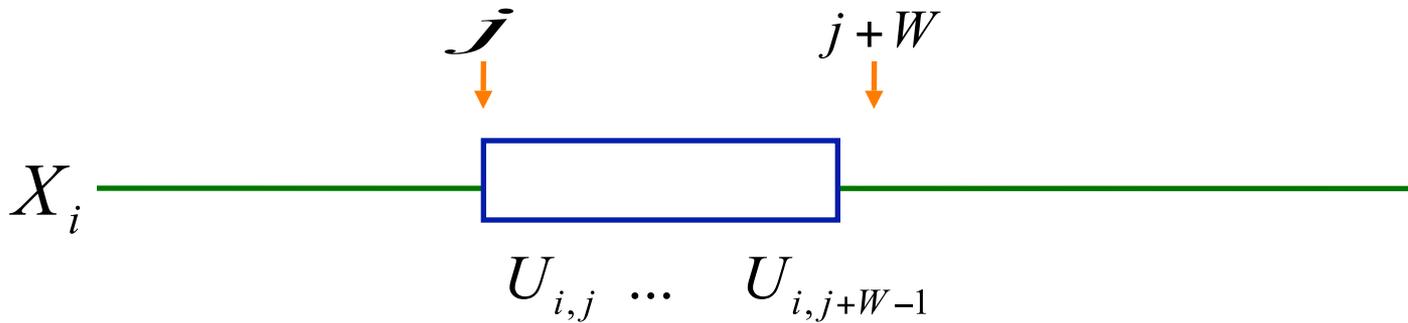
“pass” p

$$U_{i,j}^{(p)} = U_{i,j}^{(p-1)} \left(1 - \max(Z_{j-W+1}, \dots, Z_j) \right)$$



Finding Multiple Motifs

$$\Pr(V_{i,j} = 1) = \min\left(\Pr(U_{i,j} = 1), \dots, \Pr(U_{i,j+W-1} = 1)\right)$$



Starting Points in MEME

- EM is susceptible to local maxima
- for every distinct subsequence of length W in the training set
 - derive an initial p matrix from this subsequence
 - run EM for I iteration
- choose motif model (i.e. p matrix) with highest likelihood
- run EM to convergence

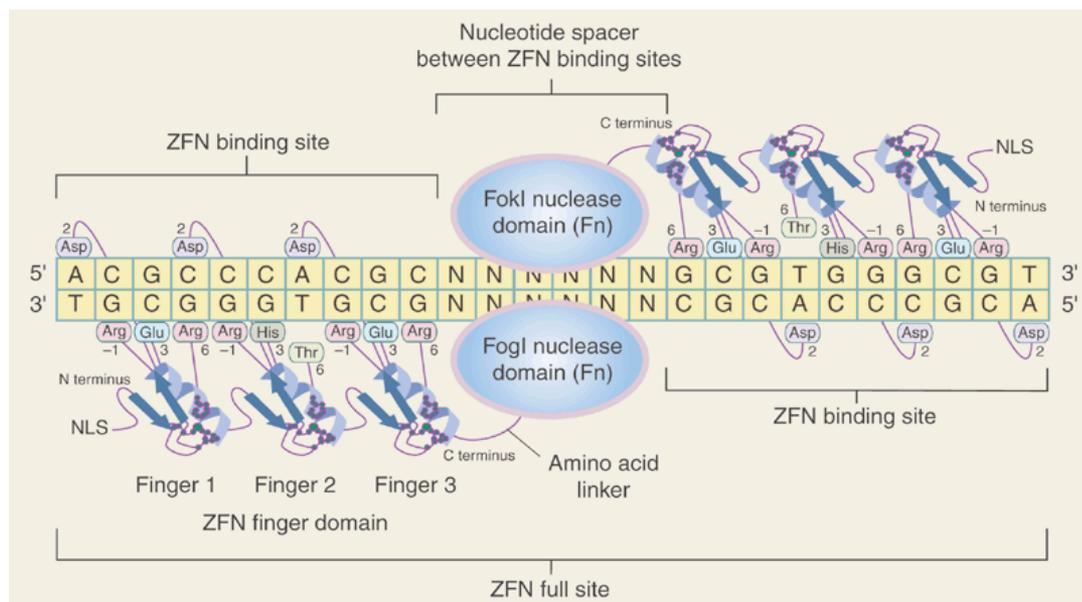
Using Subsequences as Starting Points for EM

- set values corresponding to letters in the subsequence to some value π
- set other values to $(1 - \pi)/(M - 1)$ where M is the length of the alphabet
- example: for the subsequence **TAT** with $\pi = 0.5$

		1	2	3
$p =$	A	0.17	0.5	0.17
	C	0.17	0.17	0.17
	G	0.17	0.17	0.17
	T	0.5	0.17	0.5

Using Background Knowledge to Bias the Parameters

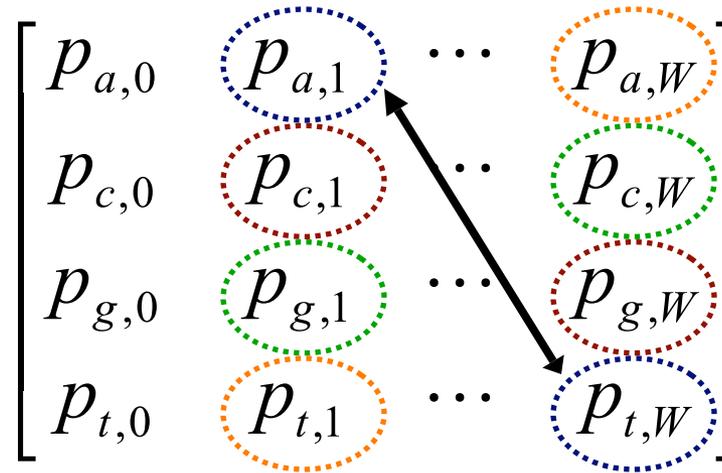
- accounting for palindromes that are common in DNA binding sites



- using Dirichlet mixture priors to account for biochemical similarity of amino acids

Representing Palindromes

- parameters in probabilistic models can be “tied” or “shared”



- during motif search, try tying parameters according to palindromic constraint; accept if it increases likelihood test (half as many parameters)

Using Dirichlet Mixture Priors

- recall that the M-step updates the parameters by:

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

- we can set the pseudocounts using a mixture of Dirichlets:

$$d_{c,k} = \sum_j \Pr(\beta^{(j)} | \mathbf{n}_k) \beta_c^{(j)}$$

Final MEME algorithm

```
procedure MEME ( $X$ :dataset of sequences )
  for  $p = 1$  to  $p_{\max}$  do
    for  $W = W_{\min}$  to  $W_{\max}$  by  $\sqrt{2}$  do
      for  $\lambda^0 = \lambda_{\min}$  to  $\lambda_{\max}$  by 2 do
        Choose good  $\theta^0$  given  $W$  and  $\lambda^0$ 
        Run EM to convergence
        Remove outer columns of motif (finetune  $W$ )
      end
    end
  end
  Output best model, adjusted for overfit
  Update priors  $U_{ij}$  for multiple motifs
end
end
```